Chapter 1

Introduction

> *"How have all those exquisite adaptations of one part of the organisation to another part, and to the conditions of life, and of one distinct organic being to another being, been perfected?"*
>
> — Charles Darwin, *On the origin of species*, 1859, p.49

1.1 Genetic and phenotypic evolution

1.1.1 Introduction

First prompted by observations made as a naturalist on board of HMS *Beagle* in South America and following more than two decades of accumulating and reflecting on facts that could shed light on the origin of species, Charles Darwin proposed natural selection on phenotypic variation as the main mechanism of evolutionary change [1]. The essence of Darwin's argument [1, 2] can be summarised as follows:

1. There is a tendency of all organisms for geometric increase in numbers. For example, "if an annual plant produced only two seeds... and their seedlings next year produced two, and so on, then in twenty years there would be a million plants".

2. Despite this tendency, the numbers of organisms on earth cannot increase geometrically - the world could not hold them; in fact, the numbers remain more or less constant. Therefore, there must be a struggle for existence: since more eggs

or seeds and young are produced than can survive and reproduce, it follows that there must be competition for survival and reproduction[1].

3. There is variation between individuals, even within single species. Individuals with variations that confer an advantage in the struggle for existence have a greater probability to survive and procreate, passing on any heritable element of such variations to their offspring. On the other hand, disadvantageous variations will tend to be eliminated.

The action of natural selection critically depends on heritability - the extent to which variation among individuals in a population is predictably transmitted to their offspring [3]. Even after the scientific community rediscovered Gregor Mendel's work on patterns of inheritance [4], much critique of Darwinism, especially in the first half of 20th century, centred around the degree to which traits are heritable [2]. Darwin strongly believed the majority of inter-individual variation to be heritable, but his evidence was anecdotal, based on observation of plant and animal breeding and on examples of characters passed on in human families [1, pp.13-14]. Today, thanks to the statistical methods of quantitative genetics, we have measured the relative importance of genes and environment for over 17,000 human traits by studying over 2 million twin pairs [5] (Figure 1.1), and also for many traits in important agronomic and natural species [3]. Artificial selection studies have shown that "almost any species will respond to selection for virtually any trait" [3, p.682], even leading to claims that human twin studies "provide compelling evidence that all human traits are heritable: not one trait had a weighted heritability estimate of zero" [5].

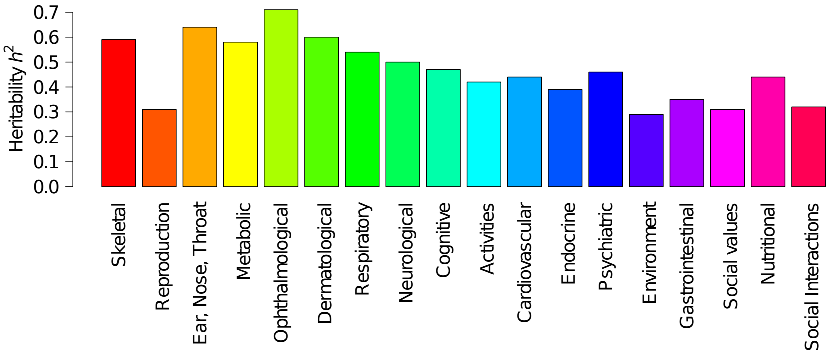


Fig. 1.1 **Heritability (h^2) estimates for over 17,000 human traits classified into 18 functional domains.** Data from [5].

---

[1] The concept of *struggle for existence* was key to Darwin's development of the theory of evolution by means of natural selection. However, some authors in do not mention it in modern presentations of natural selection, because (for example) selection can be effective even in situations when all individuals survive and reproduce.

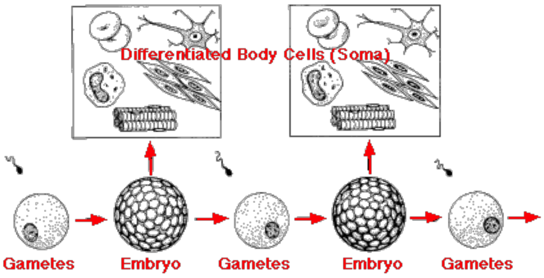


Fig. 1.2 **Separation of germline from soma.** Gametes (eggs or sperm) pass hereditary information between generations. Figure from [6].

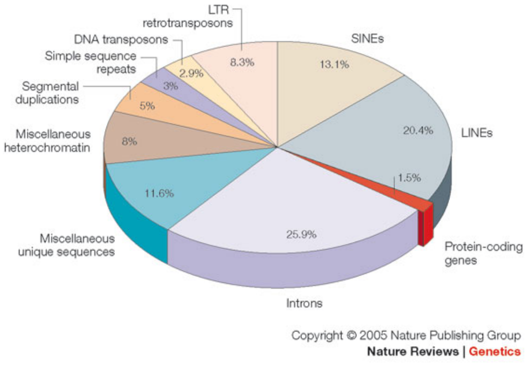
In the early 20th century, chromosomes became recognised as the physical carriers of genes between generations [7], and the fact that genes are composed of DNA first demonstrated in 1944 by Avery *et al.*, showing the ability of DNA to transform bacterial cells [8]. The role of DNA as the genetic material was confirmed in 1952 by the Hershey-Chase experiment, showing that during a bacteriophage infection only DNA is injected into the bacterium, while protein is discarded and has no further function [9]. In animals, a division is made early in development between cells that will produce eggs or sperm (the germ line) and the rest of the body (somatic cells). The German biologist Albert Weismann was the first to propose in 1893 that germ cells are the only cells that can pass genetic information between generations, and no hereditary information can pass from somatic cells to the germ line and on to the next generation (Figure 1.2) [10]. This postulate is known as the Weismann barrier and is widely accepted to apply to all vertebrates, although recent studies show that animal germ cells can in some circumstances be heritably modified by signals from somatic cells or the environment (e.g. in nematodes [11, 12, 13]).

Biological information is passed down the generations as the DNA of germ line cell lineage. Therefore, changes (mutations) in germ line DNA sequence are the ultimate source of the heritable component of phenotypic variation. The majority of germ line mutations occur as errors in DNA replication during cell division, arise at random, and can affect any part of the DNA sequence. Typically, a uniform mutation rate is considered to apply along the DNA sequence [14], although recent studies have shown that there is some minor variation in the mutation rate along the genome [15]. The genome-wide variation in germ line mutation rate is related to base composition (e.g. methylated sites with cytosine located next to guanine - CpG sites - in vertebrates have a least 10 times higher than average mutation rate), to timing of DNA replication (regions of DNA replicating late in the cell cycle tend to have a higher mutation rate), and to transcription [15].

1.1.2 Vertebrate genome architecture

Known vertebrate genome sizes vary over almost three orders of magnitude: between ~320 million base-pairs (Mb) in Green pufferfish *Tetraodon fluviatilis* and 120,000Mb in Marbled lungfish *Protopterus aethiopicus* [16]. The size of the human genome is ~3,000Mb. There appears to be with little correspondence between genome size and the external characteristics of an organism. Similarly, there is little correspondence between the number of protein-coding genes and the size, cognitive capabilities, or the number of distinct tissue types in an organism throughout eukaryotes: the number of genes in the nematode worm *Caenorhabditis elegans*, pufferfish, cichlids, and human is comparable (20,447 *C. elegans*, 18,523 in *Takifugu rubripes*, 21,437 in the Nile tilapia cichlid, and 20,300 in human; according to Ensembl annotation [17]).

Until the late 1990s, vertebrate genomes were thought to contain mainly protein coding genes but this view has changed dramatically over the last 15 years [19]. The Human Genome Project and subsequent DNA sequencing of dozens more vertebrate species revealed that protein coding sequences comprise only a small fraction of the genome, while a large proportion is taken up by mostly defunct transposable elements (~45% of the human genome; Figure 1.3). Initially viewed purely as 'junk DNA', transposon derived sequences have been shown to play a constructive role in evolution, for example the evolution of functional non-coding sequences in mammals [20]. It is now also clear that vertebrate genomes contain tens of thousands of non-coding genes. This includes small RNA genes: microRNAs, post-transcriptional regulators of gene expression [21, 22], and piRNAs, involved in repressing transposable element activity and regulating gene expression specifically in germ-line cells [23]. Thousands of long intergenic non-coding RNA genes (lincRNAs) were also identified, with diverse roles in multiple biological processes [19]. In fact the majority of the non-repetitive genome is transcribed at least in some cell types and/or has been assigned a biochemical function (such as protein binding), leading to claims by the ENCODE consortium that 80% of the genome is functional [24]. However, it is necessary to distinguish between 'selected



Copyright © 2005 Nature Publishing Group
Nature Reviews | Genetics

Fig. 1.3 **Components of vertebrate genomes.** Data for the human genome, from [18].

effect' function of a sequence (i.e. the effect for which it was selected) and 'causal function' (i.e. any function it performs)[2] [25]. Selected effect function can be inferred from evolutionary conservation: comparisons of multiple vertebrate genomes revealed that approximately 6% (3-8% depending on the estimation procedure) of the sequence is biologically functional in this sense [26], and all comparative studies reviewed by Ponting and Hardison in 2011 put the proportion of functional bases in the human genome at below 15%. In summary, the genome encodes a wide variety of genic and regulatory elements, most of vertebrate functional sequence is non-coding, and at least 85% of the genome sequence appears to be evolving without perceptible evolutionary constraint.

1.1.3 Molecular evolution

The view that most DNA nucleotide substitutions may be selectively neutral or nearly neutral became increasingly influential during the 1960s. Following the discovery that amino acid substitutions in protein sequences accumulate at a constant rate, Pauling and Zuckerkandl coined the term 'molecular clock' and proposed using DNA and protein sequences to infer phylogenetic relationships between taxa [27]. Deciphering the genetic code [28] revealed that the DNA sequence of a protein-coding gene can change in a way that does not affect its amino acid sequence (synonymous DNA substitutions), and Motoo Kimura suggested that "as functional constraint diminishes, the rate of evolution converges to that of synonymous substitutions" [29]. The 'neutral theory' of molecular evolution was developed and provided a set of baseline expectations for DNA variation and change in the absence of natural selection.

For an *ideal population* satisfying assumptions of the Wright-Fisher model (including, importantly, that "all parents have an equal expectation of being the parents of any progeny" [30, p.205], and that the population size does not vary over time), the neutral substitution rate can be derived as follows: let μ be the average mutation rate per base-pair (bp) per generation and let there be N_e breeding diploid individuals. The population generates $2N_e\mu$ mutations per bp per generation. The frequency of a new neutral allele arising due to mutation is initially $\frac{1}{2N_e}$. The allele frequency can change between generations due to random sampling ('genetic drift') and the probability of

---

[2] For example, the selected function of a heart is to pump blood, whereas causal functions may include "adding 300g to body weight, producing sounds, and preventing the pericardium from deflating onto itself" [25]

fixation of a neutral allele is equal to the allele frequency [30]. Therefore the:

$$\text{neutral substitution rate} = 2N_e\mu \times \frac{1}{2N_e} = \mu, \tag{1.1}$$

so independent of demography, and the mean time to fixation of a new mutation is $4N_e$ [30].

The *effective population size* N_e is an important factor influencing the strength of genetic drift, and also its effect on selection. The main forms of natural selection on the molecular level are listed in Table 1.1. For an allele under directional selection, the relative influence of selection and genetic drift is determined by N_e and by s the *selection coefficient*. If $N_e s \ll 1$ then the fate of the allele will be determined primarily by genetic drift. On the other hand, if $N_e s \gg 1$, it will be determined primarily by selection. For example, an allele with a selective advantage $s = 0.001$ and current frequency of 0.1 has 86.5% probability to reach fixation in a population where $N_e = 10,000$ ($N_e s = 10$), but only 10.9% probability to reach fixation in a population where $N_e = 100$ ($N_e s = 0.1$). It is clear that alleles under weak selection behave as nearly neutral in small populations, and demographic changes that affect N_e have a profound effect on genetic variation in a population.

Table 1.1 **Main forms of natural selection.** Based on [3].

Category	Selection	Description
Directional	Positive selection	Exerts force to increase the frequency of alleles that confer selective advantage
	Purifying selection	Acts to reduce the frequency of disadvantageous alleles, thus preserving functional sequences from being degraded by new mutations
Non-directional	Balancing selection	Selection acts to maintain two or more alleles at one locus, for example because heterozygous individuals have a higher fitness than homozygous ones (heterozygote advantage) or because the fitness of an allele depends on its frequency in the population

Figure 1.4 summarises the forces that influence genetic composition of natural populations. New mutations and inward migration create new *segregating sites* (polymorphic loci with two or more alleles), while directional selection and genetic drift remove variation. The balance between these forces is influenced by demographic events which change the effective population size N_e.

It is important to note that alleles on the same chromosome are not inherited independently but are physically linked to each other. The combination of alleles located together on the same physical chromosome is called a *haplotype*[3]. Linkage is

---

[3] The word *haplotype* may refer to all alleles on a chromosome or just alleles that are physically close in a particular region: e.g. a 5Mb haplotype on chromosome 1.

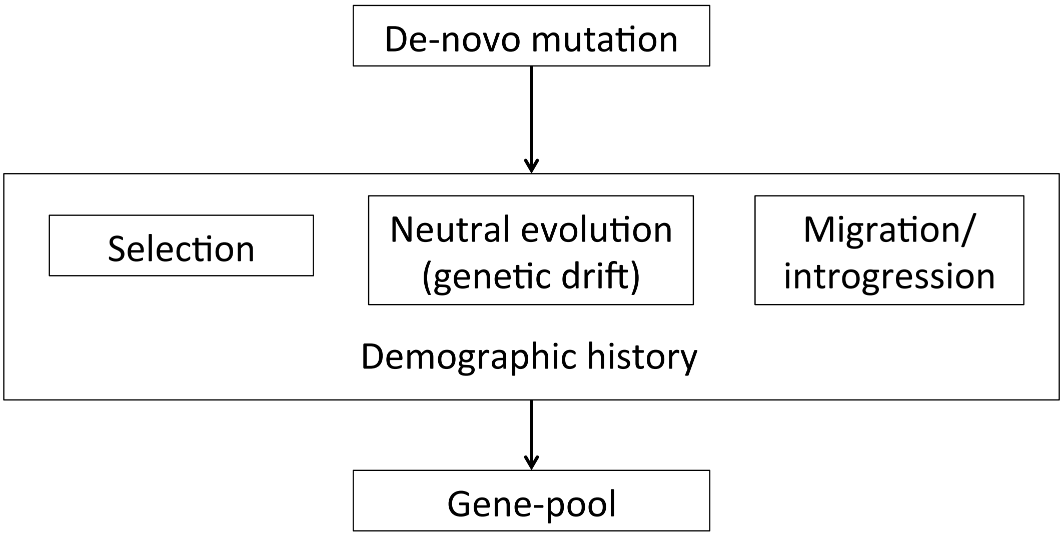


Fig. 1.4 **Forces influencing genetic composition of natural populations.**

eroded by recombination during meiosis and the probability of recombination between an allele and one of the other linked alleles on the chromosome is a monotonic function of the physical distance between them (i.e. the further the other allele, the greater the probability of recombination). Therefore, it is possible to use recombinant frequency in offspring between alleles heterozygous in the parents to detect linkage (i.e. determine if two alleles are on the same chromosome) and to generate a *genetic map* - a map describing the order of alleles along a chromosome and the distances between them in centimorgans (cM), units corresponding to 0.01 probability of recombination [3].

Consider two biallelic loci: {A,a} and {B,b}. If the relationship between the A and B alleles was random, we would expect the frequency of the haplotype AB (f_{AB}) to be the product of the frequencies of the A and B alleles (f_A and f_B):

$$f_{AB} = f_A \times f_B \tag{1.2}$$

and the two loci would be said to be at *linkage equilibrium.* Against this baseline, non-random association between alleles, referred to as *linkage disequilibrium* (LD), can be assessed. LD provides a measure of the effect of linkage on the genetic variation in a natural population. Two common measures of LD are D and r^2 [31]:

$$D = f_{AB} - f_A f_B \tag{1.3}$$

$$r^2 = \frac{D^2}{f_A(1 - f_A)f_B(1 - f_B)} \tag{1.4}$$

Linkage gives rise to the *hitchhiking effect* whereby an allele can get a lift in frequency from positive selection acting on a nearby allele on the same chromosome [32], and also to *background selection* whereby purifying selection against deleterious alleles reduces genetic variability at linked neutral sites [33]. Marked reduction of genetic variation in regions of low recombination has been observed in many organisms, including human, but the relative contributions of hitchhiking and background selection to this phenomenon are hard to ascertain [34]. However, in regions of normal recombination, the hallmark of fixation of a new strongly beneficial allele is a marked reduction in genetic diversity due to hitchhiking - a pattern referred to as *selective sweep*. Such a pattern cannot be caused by background selection in regions of normal recombination and therefore is considered a signature of recent positive selection [34].

1.1.4 Speciation

Evolution generates genetically and phenotypically distinct groups of organisms [35]. In sexually reproducing organisms, the discontinuities between the groups arise from reproductive isolation. Biological species can be considered to be units of evolution [36] and arise when reproductive isolation is complete, although limited gene exchange that does not affect the essential integrity of the species is possible [35, 36][4]. The process of speciation, i.e. the origin of species, is then a process of building up of reproductive barriers, and increasing genetic and phenotypic differentiation (Figure 1.5).

Divergence during speciation is continuous. The strength of reproductive isolation and the degree of genetic and phenotypic clustering have been shown to vary quantitatively [38]. Charles Darwin wrote: *“I look at individual differences…as of high importance to us as being the first step towards such slight varieties as are barely thought worth recording… And I look at varieties which are in any degree more permanent, as steps leading to more strongly marked and more permanent varieties; and these latter as leading to sub-species, and to species.”* [39, p.42].

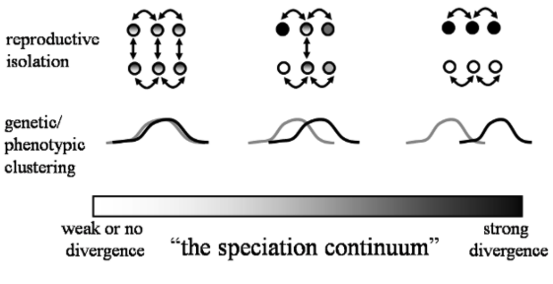


Fig. 1.5 **The speciation continuum.** Figure from [37].

---

[4] I adopt the widely accepted ‘biological species concept’. For alternative species definitions and a discussion of different species concepts see [35].

Speciation in animals typically progresses too slowly for humans to be able to directly observe it happening. We are able to measure the isolating barriers between current pairs of species and study their genetic and ecological causes. However, isolating barriers continue to evolve and accumulate after speciation is complete and the causes that initially led to speciation may be obscured by evolution post speciation. Therefore, speciation studies often focus on populations that are only partially reproductively isolated, obtaining ‘snapshot’ views of different stages along the speciation continuum. Not all partially isolated populations will go on to eventually form species, but investigation of many different taxa at many different stages all along the speciation continuum, from newly forming varieties to pairs of well defined species, will allow us to draw general conclusions about the process of speciation [35, 38].

It is useful to understand that the word *species* does not describe any intrinsic characteristics: a species can only be defined relative to other species [35]. Coyne and Orr note two common features of speciation: *epistasis*, whereby “genes that evolve in one group produce reproductive isolation by interacting with genes evolving in another group” and *pleiotropy*, whereby “characters that evolve within a group have the side effect of producing isolating barriers” [35, p.56].

In the literature, there has been a lot of focus on the geographical context of speciation [35]. A range of definitions, based either on biogeography or on population genetics, have been proposed for different geographical modes of speciation [40]. For example, speciation may be deemed *sympatric* when isolating mechanisms develop in the absence of geographic barriers to free interbreeding (i.e. individuals are within the average dispersal distance of one another [41]); *allopatric* when geographic separation causes initial split between populations, gene exchange ceases, and reproductive barriers evolve in isolation[5]; and *parapatric* when gene exchange at the onset of speciation is partially limited by geography.

1.1.5 Genome sequencing and bioinformatics

Technological advances in DNA sequencing [42, 43] since the introduction of the Sanger method in 1977 [44, 45] have led to there now being over 10^{15}bp of sequence in the ENA data depository [46], and the cost of sequencing is continuing to fall exponentially [47]. Nevertheless, significant challenges remain.

All of the currently available sequencing methods fall far short of being able to obtain the end-to-end sequence of a whole vertebrate chromosome. Instead, sequencers

---

[5] Reproductive barriers evolved in allopatry can only be observed in cases where the species come into secondary contact, or in laboratory experiments

output large numbers of short reads that contain errors and need to be analysed computationally in order to provide useful information. Illumina, currently the most cost effective sequencing platform [47], produces ~100bp reads. To obtain a new genome sequence for a species, the reads need to be joined up through overlaps into a continuous sequence. Such *de novo* genome assembly is a task analogous building a jigsaw puzzle consisting of hundreds of millions or even billions of pieces, and is one of "the most complex computations in all of biology" [48]. To obtain sufficient overlaps between reads and also to be able to recognise errors in individual reads, every site in the genome needs to be sequenced multiple times (typical average *genome coverage* required for *de novo* assembly of 100bp reads is 40-100×). On the other hand, when an assembled *reference genome* is already available, it is possible to align to this reference reads from other individuals of the same species, or of a closely related species, and to infer genetic variation in the form of differences between the mapped reads and the reference [49]. The alignment approach is substantially cheaper and faster because it typically requires lower genome coverage (~5-20×) and is much less computationally intensive than *de novo* assembly [43].

1.2 Evolutionary and speciation genomics

The search for the molecular basis of evolutionary processes leading to adaptation and organismal diversity predates DNA sequencing. In one of the earliest and best known studies, V. M. Ingram, working at the Cavendish Laboratory of the University of Cambridge, discovered in 1957 the single amino acid difference between normal and sickle cell haemoglobin [50]. The sickle cell allele confers a large survival advantage to heterozygous individuals in regions with a high incidence of malaria, but is highly deleterious and often lethal when homozygous [3]. Therefore, the allele is under balancing selection in regions with high incidence of malaria, but is under purifying selection in other parts of the world. Other pre-genomic studies focussed on differences between the rates of synonymous and non-synonymous nucleotide substitutions in protein-coding genes. McDonald and Kreitman in 1991 compared within-species to between-species polymorphisms in the alcohol dehydrogenase *Adh* gene of three *Drosophila* fruit fly species. They found an excess of non-synonymous changes in between-species comparisons - evidence for repeated fixation of advantageous alleles and long-term adaptive evolution in this gene [51]. This is consistent with the role of *Adh* in alcohol tolerance and utilisation and adaptation of *Drosophila* to new feeding niches involving fermenting fruit [52].

While the above and other similar studies provided fascinating insights, they were limited to single genes. The availability of data from many genes and, later, whole genomes enabled scientists to draw more general conclusions about adaptive evolution. For example, a comparison of 43 genes between *Drosophila yakuba* and *D. simulans* suggested that “45% of all amino-acid substitutions have been fixed by natural selection, and that on average one adaptive substitution occurs every 45 years in these species” [53]. The comparison of the human and mouse genomes revealed that ~80% of mouse genes have clear 1:1 *orthologs* in human, originating from common ancestral sequence [54], but for example the *V1r* family of pheromone receptors has ~160 functional genes in mice and only 5 in human [3], consistent with the key role of pheromones in social and reproductive behaviour in mice [55]. Finally, comparisons of human and chimpanzee genomes provided a virtually complete catalog of genetic differences between our species and our closest relatives, revealing ~35 million single nucleotide substitutions, ~5 million small insertions and deletions, differential signatures of transposable element activity, and a number of larger chromosomal rearrangements [56]. In 2006, Katherine Pollard and her colleagues compared 17 available reference genomes and found 49 ‘human accelerated regions’ (HARs) with significantly accelerated rate of substitution in the human lineage, but strong sequence conservation across reptiles, birds, and other mammals [57]. Many HARs are located near genes involved in neurodevelopment, and the top HAR contains a long non-coding RNA expressed in the developing neocortex [57]. 96% of HARs are not in protein-coding genes, and broader comparisons of vertebrate genomes also suggest that functional non-coding sequences are more abundant and tend to evolve faster than protein coding sequences [19]. This evidence points to changes in gene regulatory regions, rather than differences in genes themselves, being primary genetic drivers of adaptive differences between closely related species. Overall, all the above examples illustrate the power of comparative genomics to shed light on the molecular basis of organismal diversity.

Population genetic data (i.e. patterns of genetic variation within populations) are informative about more recent selection. A variety of methods have been developed to infer positive selection by comparing observed population genetics data with expectations under neutrality [58]. As noted in section 1.1.3, a hallmark of a recent selective sweep is a marked reduction in genetic diversity in the genomic region surrounding the beneficial allele (Figure 1.6a). The process is accompanied by changes in the population frequencies of segregating alleles in the region (Figure 1.6b). As the beneficial allele rises in frequency, derived (i.e. arising via recent mutations) alleles on the same haplotype also rise to high frequencies and the excess of high-frequency derived alleles can be detected using Fay & Wu's H statistic [59]. Then, after fixation of the beneficial allele, diversity starts returning in the form of new mutations, all of which are initially at low frequencies and the surplus of rare alleles is detected using the Tajima's D statistic [60].

As the haplotype with a beneficial allele rises in frequency it creates a local distortion in the patterns of LD. A high-frequency long haplotype may be a sign of its rapid rise in prevalence, as recombination has not had enough time to break the LD and shorten the haplotype. The extended haplotype homozygosity (EHH) statistic has been designed to capture such pattern of locally elevated LD (Figure 1.6c).

Finally, if positive selection acts on a locus in one population but not in another related population, it may result in differences in allele frequencies between the two populations (Figure 1.6c). Measures of population divergence, such as F_{ST}, d_f, and d_{XY} will be discussed further in section 1.2.1.

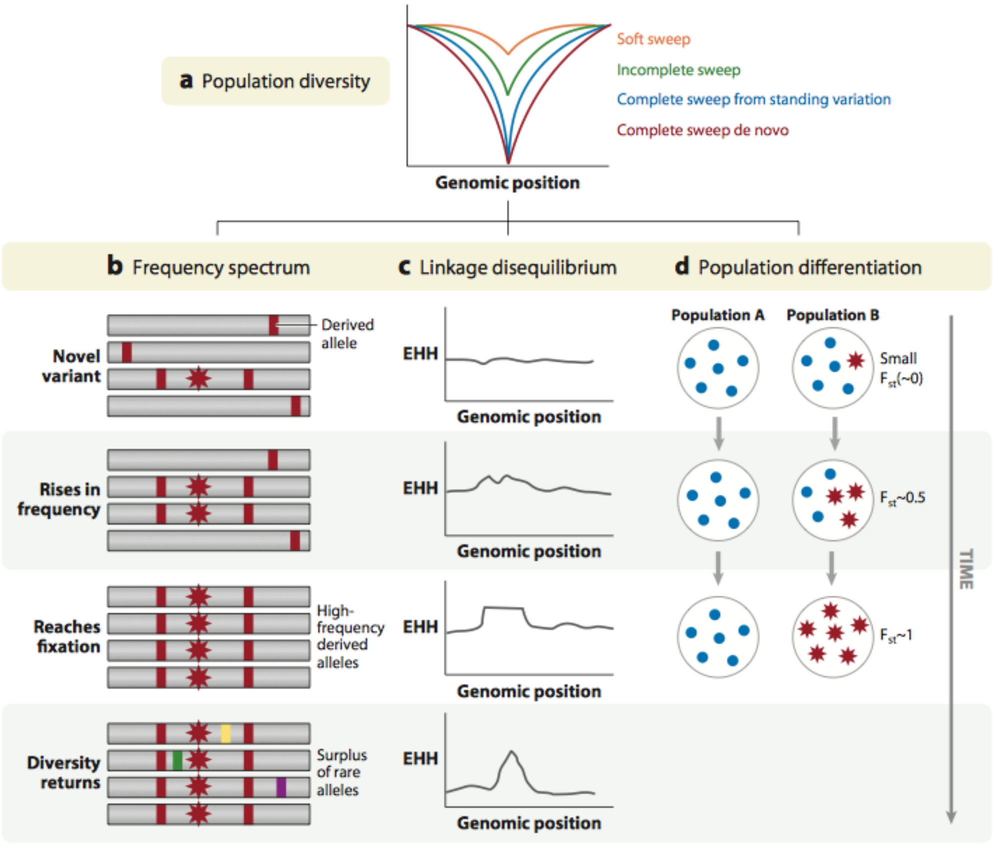


Fig. 1.6 **Signatures of selective sweeps in population genetics data. (a)** A local reduction in sequence diversity is the hallmark of all types of selective sweeps. **(b-d)** Changes in allele frequencies, local patterns of LD, and divergence between populations under different selective pressures can be used to detect positive selection in population genetic data. Figure from [58].

Population genetics is typically thought of as being focussed only on variation within a group of individuals of the same species [3]. This view is based on the assumption that genealogies of sequences sampled from a particular locus in multiple individuals *coalesce* (have a common ancestor) within the species and within-species variation has little impact on inferences concerning between-species divergence. This model (Figure 1.7A) relies on "all branches in the species tree being very long compared to within-species coalescence times" [61, p.871], and on complete isolation between species. When assumptions of hard split with long intraspecific branches are severely violated, as in Figure 1.7B, by *incomplete lineage sorting*, i.e. lineages not coalescing within the duration of their species, and/or by *introgression*, i.e. gene flow between species, the population genetics framework and methods are also appropriate to use for multi-species datasets.

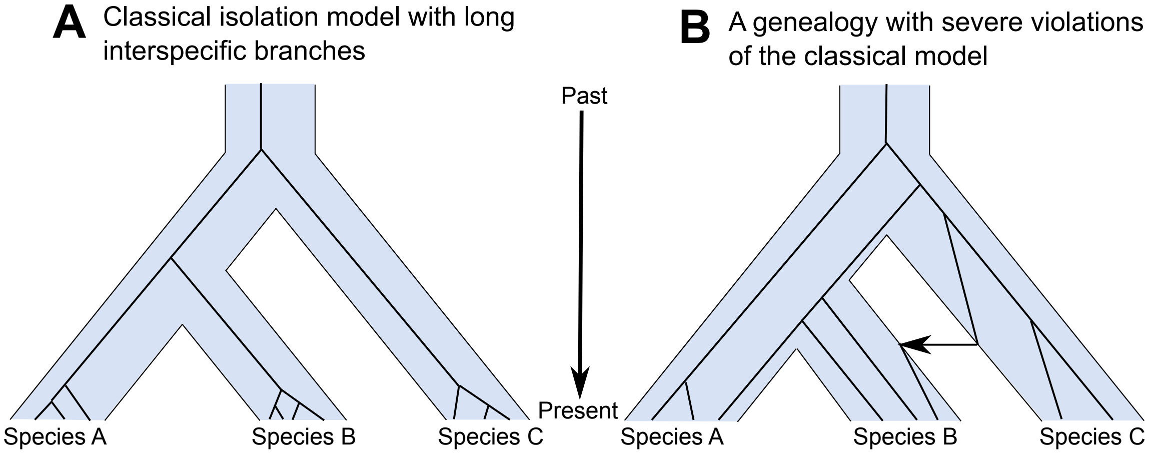

A
Classical isolation model with long interspecific branches
B
A genealogy with severe violations of the classical model
Past
Present
Species A
Species B
Species C
Species A
Species B
Species C

Fig. 1.7 **Sequence genealogies and speciation. (A)** A classical model of speciation with hard split between species, full isolation, and within-species coalescence times much shorter than between-species branches. **(B)** An example of sequence genealogy severely violating assumptions of the classical model due to pervasive incomplete lineage sorting and gene flow from species C to species B.

One of the central goals in the emergent field of speciation genomics is describing genome-wide patterns of divergence at different stages along the speciation continuum and understanding: 1) how are these patterns affected by the geographic mode of speciation (allopatric ↔ sympatric); 2) how much of the divergence at different stages is driven by selection and how much by neutral processes; and 3) the degree to which genomic signatures are common between different organisms (e.g. are the patterns of divergence along the speciation continuum in wild mice similar to the patterns in cichlid fishes?) [62].

Theory predicts a fundamental difference in genomic signatures of allopatric divergence, where reproductive isolation can in principle be simply a result of a prolonged period of genetic drift, and signatures of sympatric or parapatric speciation, where geography permits at least some continuing gene-flow. For speciation to occur in the sympatric or parapatric case, the homogenising effect of gene-flow needs to be counteracted by divergent selection. Moreover, the accumulation of loci contributing to reproductive isolation in the face of gene-flow is counteracted by recombination. Linkage between genomic loci underlying different isolating barriers mitigates the effects of recombination, and so is thought to be conductive to speciation in these scenarios. This has been demonstrated for example in flycatchers [63] and in stickleback [64]. Felsenstein showed that linkage between loci underlying prezygotic (e.g. mate choice) isolation barriers with loci underlying postzygotic (e.g. low survival of hybrids) barriers can be especially important in facilitating speciation with gene-flow [65].

1.2.1 Measures of sequence divergence

When sampling the genomes of multiple individuals from each population or species, there are multiple ways to measure their level of divergence. The currently most commonly used measures have been reviewed by Cruickshank and Hahn [66]. It is useful to distinguish relative measures (e.g. F_{ST}, d_f) and absolute measures (d_{XY}). A brief description of these statistics is provided in Table 1.2. The distinguishing feature of relative measures of differentiation is that they are influenced by within-population levels of variation. For example, a decrease in variation in population `A` due to a selective sweep would cause a rise in F_{ST} between populations `A` and `B`, but the absolute divergence between `A` and `B` would remain constant.

Table 1.2 **Commonly used measures of population divergence.** Adapted from [66].

	Measure	Description
Relative measures	F_{ST}	Normalised measure of allele frequency differences between populations
	d_f	Number of fixed differences between populations or species
Absolute measure	d_{XY}	Average number of pairwise differences between sequences from two populations, excluding all comparisons between sequences within populations.

1.2.2 'Islands of speciation' or 'incidental islands'?

A pattern with well defined genomic regions of markedly elevated F_{ST} divergence between incipient species has been observed in a large number of studies of speciation with gene-flow [e.g. 67, 68, 69, 70]. In these studies, which include cases where gene-flow occurred due to divergence in sympatry or due to secondary contact after a period of allopatry, 'islands' of high differentiation have been interpreted as loci resistant to gene-flow. These empirical observations have been accompanied by theoretical models of speciation with gene-flow, with divergent selection generating genomic 'islands of speciation', while the rest of the genome is homogenised by gene-flow [e.g. 71, 72]. Under these models, highly diverged regions (HDRs) have lower effective migration rates and harbour alleles underlying isolating traits between the incipient species - i.e. reproductive isolating genes generating 'islands of speciation'.

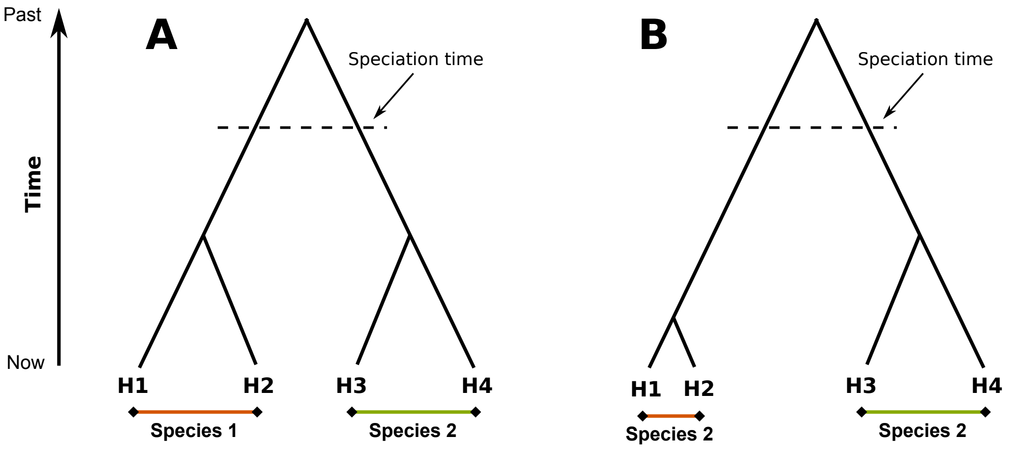
Others have offered alternative explanations [66, 73, 74], highlighting that 'islands' of high relative divergence, as measured for example by F_{ST} and d_f (described in Section 1.2.1) can also be caused by other forces in the absence of gene-flow. One alternative model that has received much attention suggests that post-split selection (e.g. a selective sweep) generates regions of reduced genetic diversity in one or both of the newly formed species, thus increasing F_{ST} in the vicinity of the selected locus [66, 73, 75]. Under this model, 'islands' of high relative divergence may be involved in post-split adaptation, or specialisation, or may reflect background selection present in any organism [66]. They do not directly cause speciation, and, therefore, have been referred to as 'incidental islands' [74].

The two contrasting hypotheses with regards to the origin of 'islands' of high differentiation are summarised in Table 1.3. A crucial prediction of the 'islands of speciation' model is that that both relative and absolute measures of divergence should be high in regions resistant to gene-flow, whereas the 'incidental islands' model predicts that absolute divergence will not be affected [66, 73]. Mindful of this prediction, Cruickshank and Hahn [66] last year re-analysed published sequence data for eight recently diverged taxa with putative 'islands of speciation' and showed that, in all cases, absolute divergence (d_{XY}) in the islands was lower than outside - not consistent with the 'islands of speciation' hypothesis. It is therefore likely that many previously reported genomic islands do not represent the first steps in the formation of new species. The question is not settled, however, with prominent evolutionary biologists defending the role of the divergent islands in speciation [76], citing for example a recent study of genomic divergence in the face of gene-flow between the European hooded and carrion crows. This study revealed a single prominent island of high F_{ST} and d_f that exhibits low absolute divergence but nevertheless is very likely involved in the maintenance of reproductive isolation [77].

Table 1.3 **Interpreting genomic 'islands' of high differentiation.** Two alternative hypotheses are introduced, together with their predictions for levels of relative (F_{ST}, d_f) and absolute (d_{XY}) differentiation in the islands. Based on [66, 74].

Model	Cause	Interpretation	Predictions
Islands of speciation	Lower effective migration	Loci underlying reproductive isolating barriers/traits thereby causing speciation	High F_{ST}, d_f High d_{XY}
Incidental islands	Lower effective population size	Loci involved in post-split adaptation or simply background selection; alleles within islands do not cause speciation	High F_{ST}, d_f Low d_{XY}

To better understand why a selective sweep in a newly formed species does not affect d_{XY}, consider the sequence genealogies shown in Figure 1.8. Each genealogy connects haplotypes from the same individuals belonging from two species (haplotypes H1, H2 from species 1 and haplotypes H3, H4 from species 2); the (A) panel shows the genealogy at a neutral locus, whereas the (B) panel shows a locus under recent (post-speciation) positive selection in species 1, reflected in the more recent common ancestor of haplotypes H1 and H2. Despite the differences between the genealogies, the average distance (H1,H2 $\longleftrightarrow$ H3,H4) is exactly the same between the two panels. Thus, d_{XY} remains unchanged.



A neutral locus

Post-speciation selection in species 1

Fig. 1.8 **Effect of different sequence genealogies on F_{ST} and d_XY.** The figure shows the genealogies of four haplotypes from two species (two haplotypes from species 1 and two haplotypes from species 2) at **(A)** a neutral locus and **(B)** a locus under recent positive selection (selective sweep) in species 1. Figure adapted from [66].

1.3 East African cichlid radiation

1.3.1 Cichlid fish

Cichlids (*Cichlidae*) are one of 448 families of modern bony fish (*Teleostei*) and belong to the *Labroidei* suborder [78]. The geographical distribution of cichlids and phylogenetic relationships suggest that the cichlid family originated on the Gondwana supercontinent about 150 million years ago [79]. Subsequent fractioning of Gondwana means that, today, cichlids are found in Southern India, on Madagascar, across Sub-Saharan Africa and along the Nile, and in the tropics of the Americas. Although cichlid morphology is incredibly diverse, all have a common body plan in terms of the position of the fins and the arrangement of parts of the jaws (Figure 1.9).

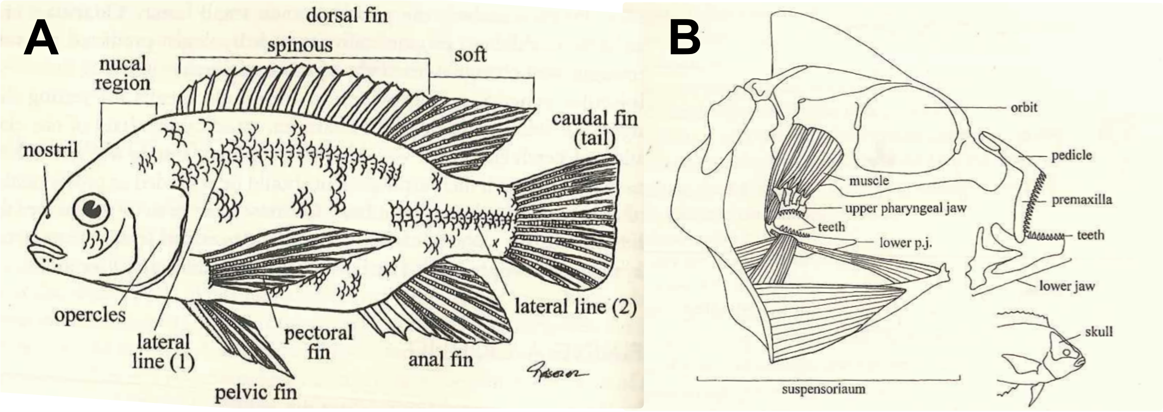


Fig. 1.9 **Major physical features of cichlids.** **(A)** The basic body plan showing the position of fins and two other distinctive features: a single nostril and an interrupted lateral line. **(B)** A cross section through the head of a cichlid, showing the unique arrangement of pharyngeal jaws. Both figures from [78].

Characteristic features of cichlids include having only one nostril on each side of the snout (fish typically have two), having a lateral line that is interrupted about two thirds of the way to the tail (fish typically have a single uninterrupted lateral line on each side), and, crucially, a unique arrangement of pharyngeal jaws.In most fishes, the upper pharyngeal jaw floats freely and the lower pharyngeal bones are split and limited to simple rocking action. Therefore, their utility for chewing food is limited. On the other hand, the upper pharyngeal jaw in cichlids has developed a protruding connection with the base of the skull and no longer floats freely. The lower pharyngeal bones in cichlids are fused into a single jaw (in common with other fishes of the suborder *Labrodei*) and developed bony outgrowths for attachment of more muscles, enabling it to produce a much greater force. Having two sets of jaws capable of chewing is believed to facilitate rapid evolution of new feeding specialisations and to contribute to the great evolutionary success of cichlids and other Labrodei [78, 80].

1.3.2 East African cichlids

Cichlids are the most species rich and diverse family of vertebrates [81]. They have undergone repeated adaptive radiations in more than 30 African lakes, although 120 more lakes across the continent have been colonised by cichlids without speciating [82]. Lake depth and sexual dichromatism (colour differences between males and females) are significantly associated with African cichlid radiations [82]. There are approximately 2,000 distinct cichlid species in East Africa, with great diversity in habitat, behaviour, feeding apparatus, craniofacial morphology, pigmentation, and body shapes [83]. The diversity of tooth and jaw morphologies and their association with habitats and ecological niches is evidence for the importance of these traits in the East African radiations [83] (Figure 1.10).

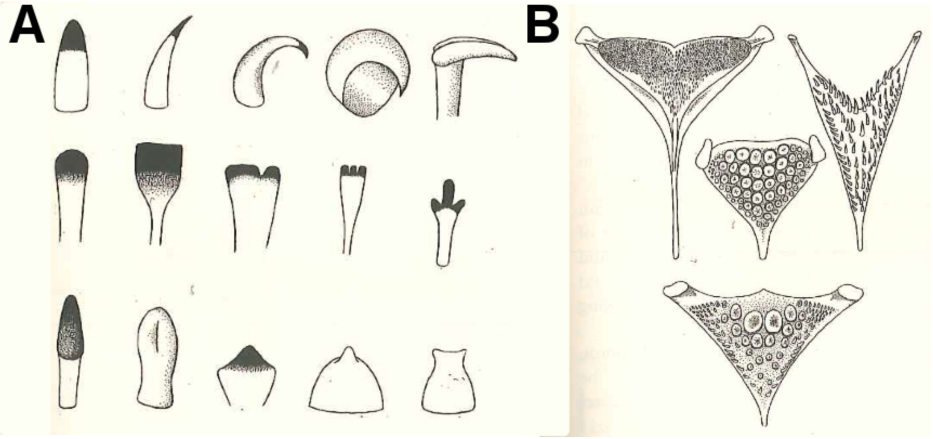


Fig. 1.10 **Evolution of cichlid feeding apparatus in East Africa.** **(A)** Outer jaw teeth of several species of African cichlids. Teeth with cusps and a flat top are common in algae scrapers, while pointy conical teeth are common in predators. **(B)** Lower pharyngeal jaws from an algae feeder (top left), piscivore (top right), a specialised molluscivore (middle) and a more generalist molluscivore (bottom). Figures from [78].

Nowhere have cichlids radiated faster and more spectacularly than in the Great Lakes of East Africa: Lakes Malawi, Tanganyika, and Victoria. The lakes contain hundreds of species, virtually all of which are endemic to each. The ~200 species rich Lake Tanganyika radiation is believed to have been seeded by eight independent cichlid lineages and, being 9-12 million years old, it hosts the oldest and most diverse assemblage of cichlid fishes [84]. Both the smallest and the largest known cichlid species are endemic to Tanganyika: *Neolamprologus brevis* at 3 cm and *Boulengerochromis* *microlepis* at 80 cm. Tanganyikan cichlids are grouped into 12 tribes, one of which, the tribe *Haplochromini*, contains two endemic species and several lake-river generalists found throughout the Tanganyikan catchment area.

The cichlid radiations in Lakes Malawi and Victoria have been seeded by Haplochromini riverine lineages that shared a common ancestor with the Tanganyikan haplochromines approximately two million years ago. It is thought that ancestral haplochromines evolved in Lake Tanganyika, colonised river systems of East Africa and through the rivers also entered the newly emerging Lakes Malawi and Victoria [85] (Figure 1.11). Therefore, while there are only a handful of Haplochromini cichlids in Lake Tanganyika, almost all of the over 1,000 species of Lakes Malawi and Victoria are haplochromines [85].

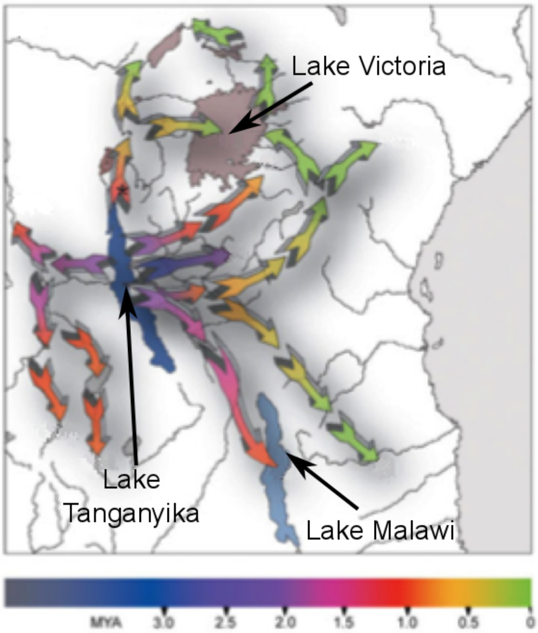


Fig. 1.11 **The 'out of Tanganyika' model of East African cichlid radiations.** The ancestors of today's haplochromines used in Lake Tanganyika and used existing and ancient river systems to spread across East Africa, also seeding the spectacular radiations of Lake Malawi and Lake Victoria. Figure adapted from [85].

The ~1-2 million years old cichlid flock of Lake Malawi has most likely been seeded by fish similar and ancestral to the lake-river generalist *Astatotilapia calliptera*, possibly in multiple repeated colonisation events over hundreds of thousands of years [86]. Most Lake Malawi cichlids can be assigned to two major lineages (each with ~250 species): the rock-dwelling 'mbuna', and the 'sand-dwellers'. There are also two genera of open-water predators, each with ~10-20 species: *Diplotaxodon* and *Rhampochromis* [87].

Lake Victoria, the youngest of the three Great Lakes, dried up some 15,000 years ago. This event created a bottleneck in the cichlid population of the Lake Victoria basin. One study suggests that a 30-50 fold decline in cichlid populations ensued, with surviving cichlids occupying refugia in rivers and other lakes in the area [88]. These haplochromine cichlids began to re-enter the re-emerging Lake Victoria from multiple directions (Figure 1.11) some 12,000 years ago and formed a genetically diverse founding population. Therefore, even though the Lake Victoria species flock is only thousands of years old, the most common ancestor of these fish can only be traced back to the time Haplochromini left Lake Tanganyika [88]. The ancestral genetic variants present in the fish re-entering the lake 12,000 years ago have since been 'mixing-up' through hybridisation and new species are emerging at an astonishing rate, adapting to conditions in different parts of the lake (e.g. ref [89]).

Frequent occurrence of independent evolution of similar phenotypes in different lakes and lineages suggests an important role of natural selection in generating these radiations [90, 91]. Examples of ecologically driven adaptations include pelagic zooplanktivores with a large number of gill rakers, rock-dwelling algae scrapers with many rows of fine brush-like teeth, snail crushers with hypertrophied pharyngeal jaws, several groups of dark-adapted species with greatly enlarged lateral-line sensory apparatus, large pelagic piscivores with a single row of unicuspid teeth, and insect-eaters with 'fat' lips used to seal crevices in the rocks and prevent prey from escaping. These populations of cichlid 'variants' present an unparalleled opportunity to study the genetic basis of how complex traits evolve in vertebrates. This includes traits occurring in the later stages of the life cycle, not included in many traditional induced mutagenesis screens.

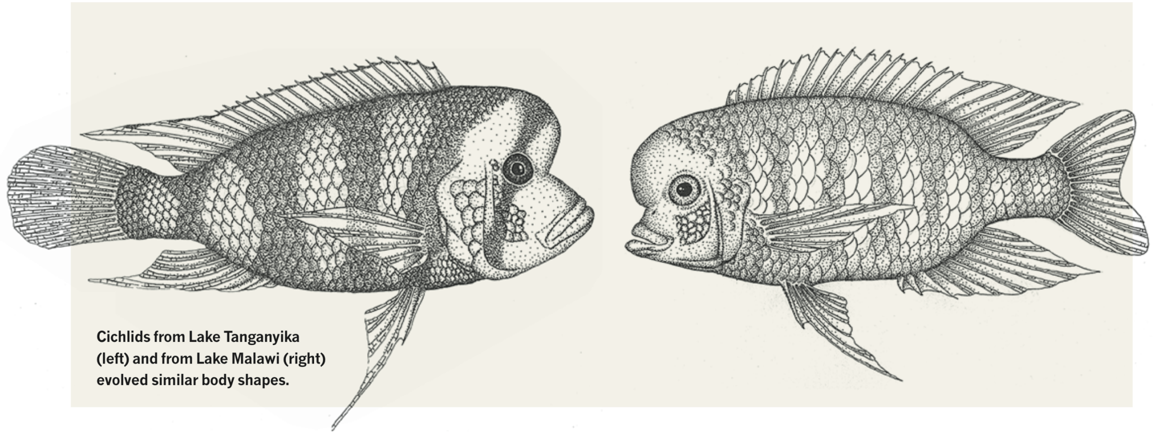

Cichlids from Lake Tanganyika (left) and from Lake Malawi (right) evolved similar body shapes.

Fig. 1.12 **Parallel evolution of body shapes in Lakes Malawi and Tanganyika.** Figure from [92].

1.4 Overview of the remainder of this thesis

This thesis describes some of the first steps in the application of whole genome sequence data to study the exceptional diversity of East African cichlids. The rest of the thesis is organised in six chapters:

- Chapter 2 introduces reference genomes and annotations generated by the Cichlid Genome Consortium, and describes my contribution: microRNA gene annotation. This work has been published in Brawand, Wagner, Li, Malinsky *et al.*, 2014 [93]. In addition, the chapter contains unpublished work on evolution of target genes regulated by microRNAs.
- Chapter 3 provides an overview of the new whole genome data generated during my PhD and describes the bioinformatics groundwork I have done to facilitate the use of this data for scientific research on East African cichlid evolution.
- Chapter 4 takes a short detour into computational biology and describes a new method for reducing problems caused by heterozygosity during genome assembly called `trio-sga`. The method is applied to generate improved *de novo* genome assemblies from cichlid data, and its benefits are demonstrated more dramatically using the highly heterozygous *Heliconius* butterfly data. A manuscript on the `trio-sga` method is in preparation.
- Chapter 5 contains the initial analysis of Lake Malawi cichlid samples, focussing on characterisation of genomic diversity, relationships between species, and the extent of introgressive hybridisation. This will be extended by adding further data from Summer 2015 and additional analysis to form the basis of a future publication.
- Chapter 6 presents a detailed characterisation of early-stage adaptive divergence of two cichlid fish ecomorphs in Lake Massoko, a small (700m diameter) isolated crater lake in Tanzania. The work described in this chapter has been submitted as a ‘Report’ to the journal *Science* in July 2015. The manuscript received favourable reviews and has been upgraded to a full ‘Research Article’. The revised article has been submitted as Malinsky, Challis, Tyers *et al.*, ‘Genomic Islands of Speciation Separate Cichlid Ecomorphs in an East African Crater Lake’ on 11th September 2015.
- In chapter 7, I first summarise the main contributions and findings of the research presented in this thesis of and then discuss ongoing work and future directions.